CS-E5740

Complex Networks

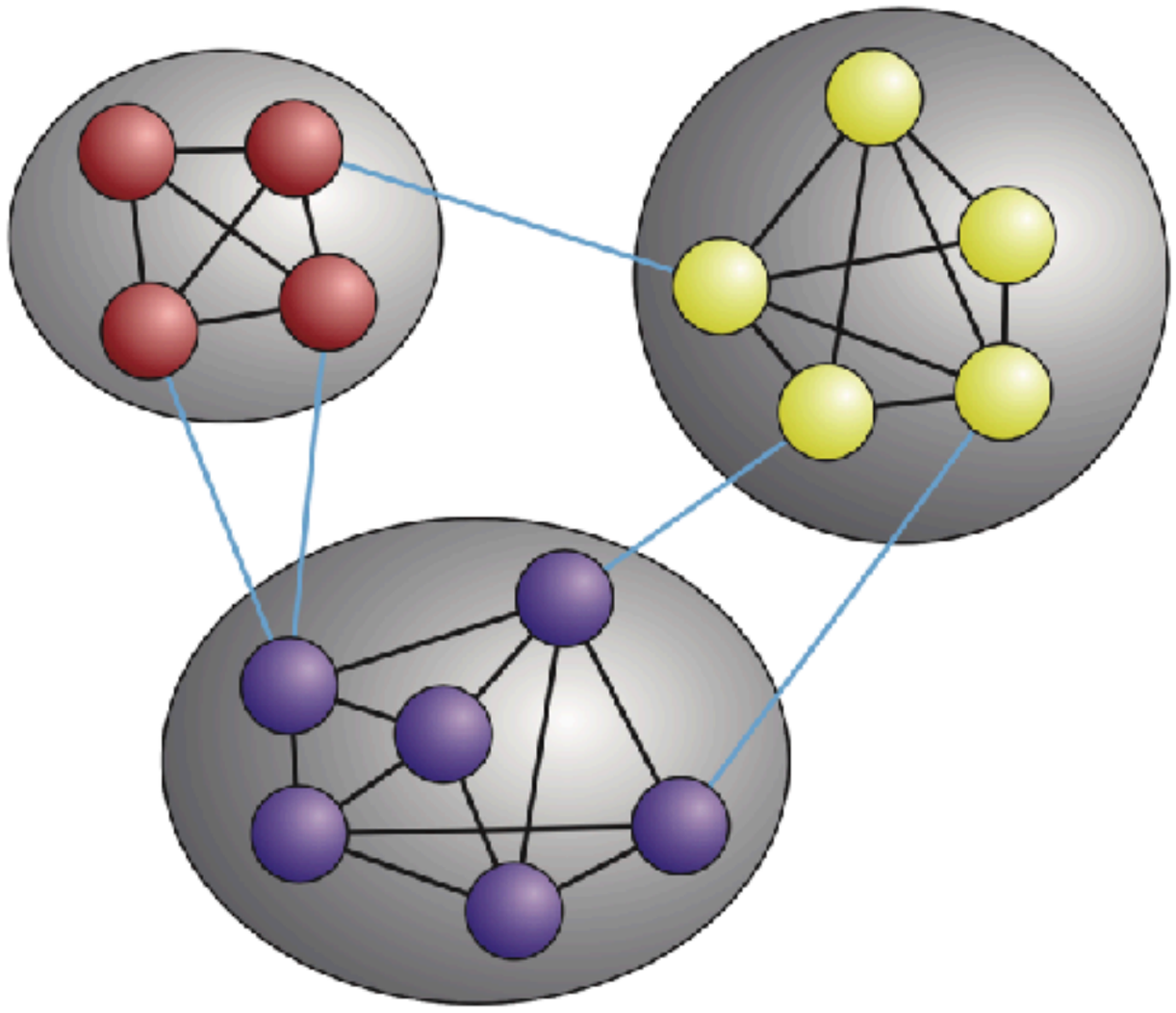
Clustering (and inference)

Course outline

1. Introduction (motivation, definitions, etc.)
2. Static network models: random and small-world networks
3. Growing network models: scale-free networks
4. Percolation, error & attack tolerance of networks, epidemic models
5. Network analysis: key measures and characteristics
6. Social networks & (socio)dynamic models
7. Weighted networks
8. Clustering, sampling, inference
9. Temporal networks & multilayer networks

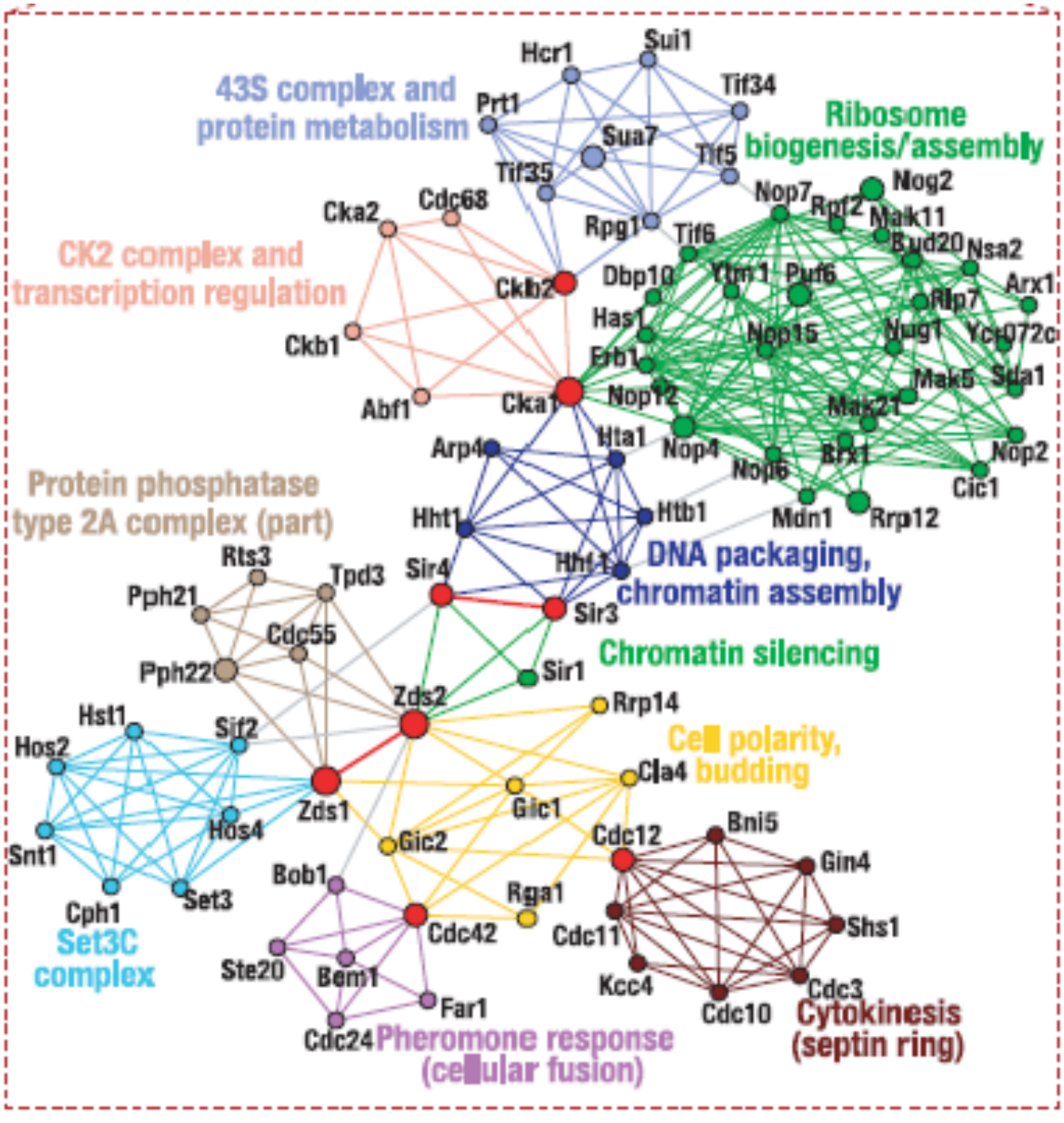
Clustering
(=community detection)

Communities, clusters, modules



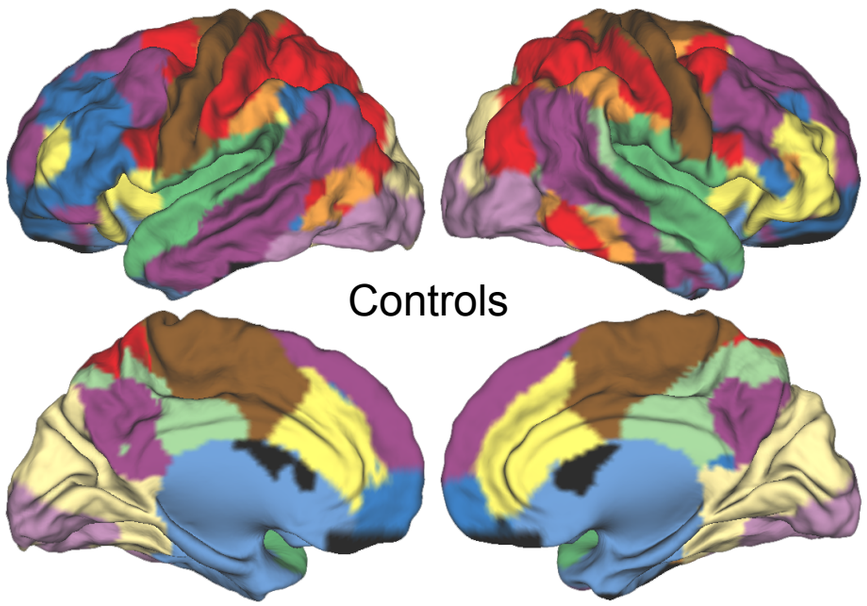
- Sets of densely connected nodes, joined by small number of links
- Ubiquitous in real-world networks
- Large number of methods for detecting them exist

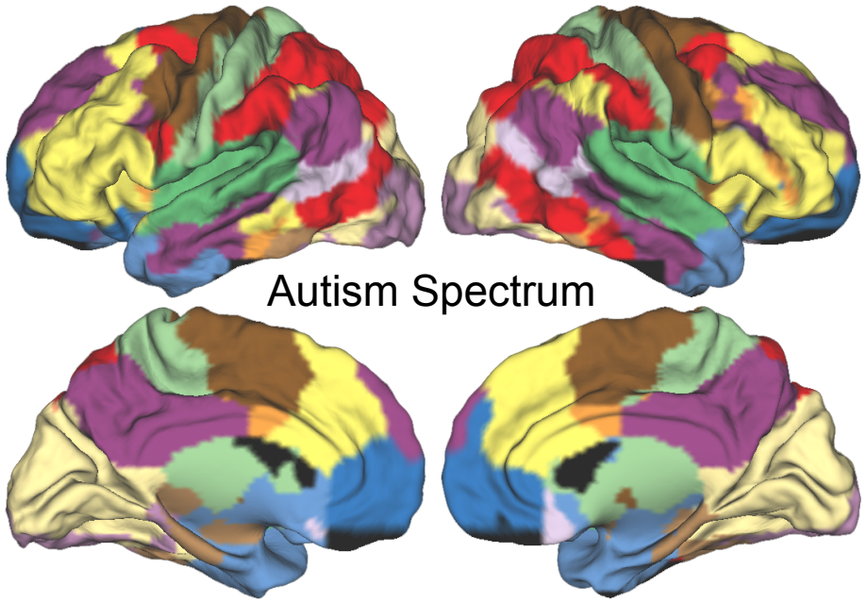
Biology: clusters are functional modules



- Proteins/genes in same cluster often participate in the same function in the cell
  - Suggests functions for unknown nodes
- Identify functional units in less studied organisms

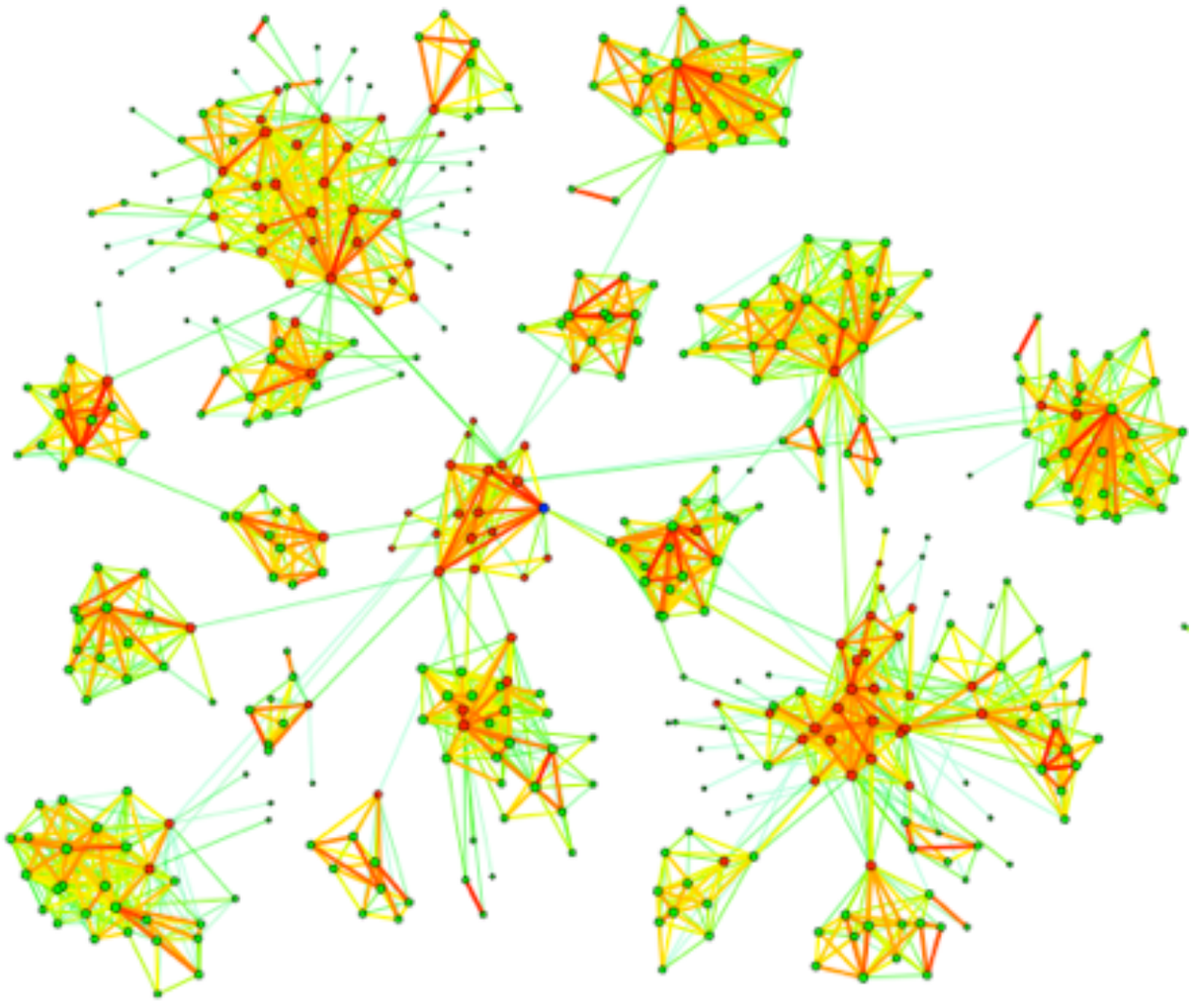
Functional brain networks: areas co-activating under given task





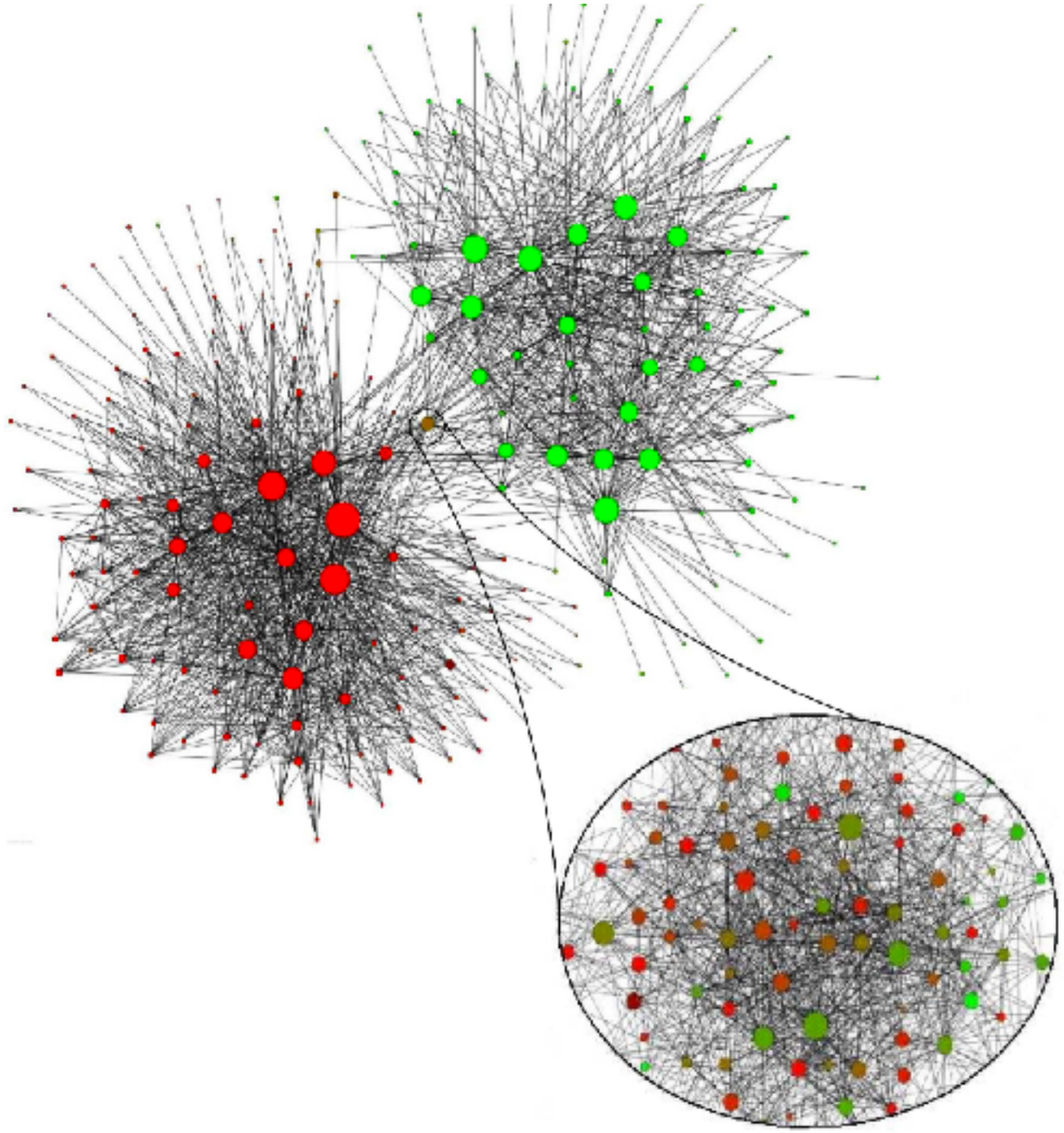
- Clusters of networks constructed of brain activity can correspond to functions in the brain
- Functional modules might differ when different tasks are performed
- Example: Compare how the functions differ in normal and autistic brain

Social networks: clusters are social groups



- It is natural for people to form social groups
- Network clustering can be used to find these "communities"
- Link weights are important in the cluster structure of social networks
  - Granovetter hypothesis

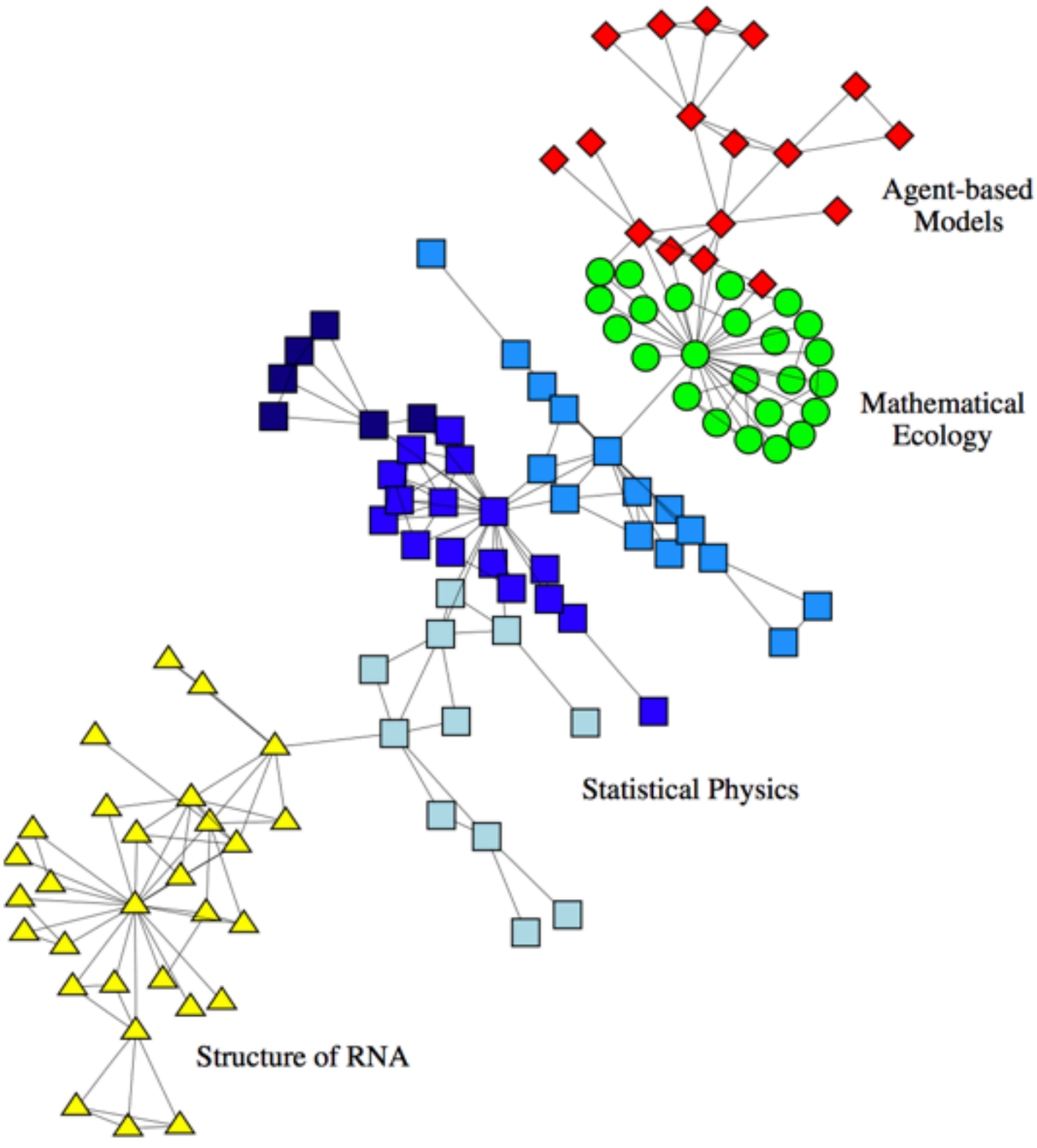
Social networks: groups at many levels



Blondel et al, J. Stat. Mech. (2008)

- Clusters in social networks can be due to homophily, geographic constraints etc.
- Cluster structure could be used to predict attributes of nodes in social networks when some nodes have missing data
  - But there are many factors: no guarantee that this works for any particular attribute

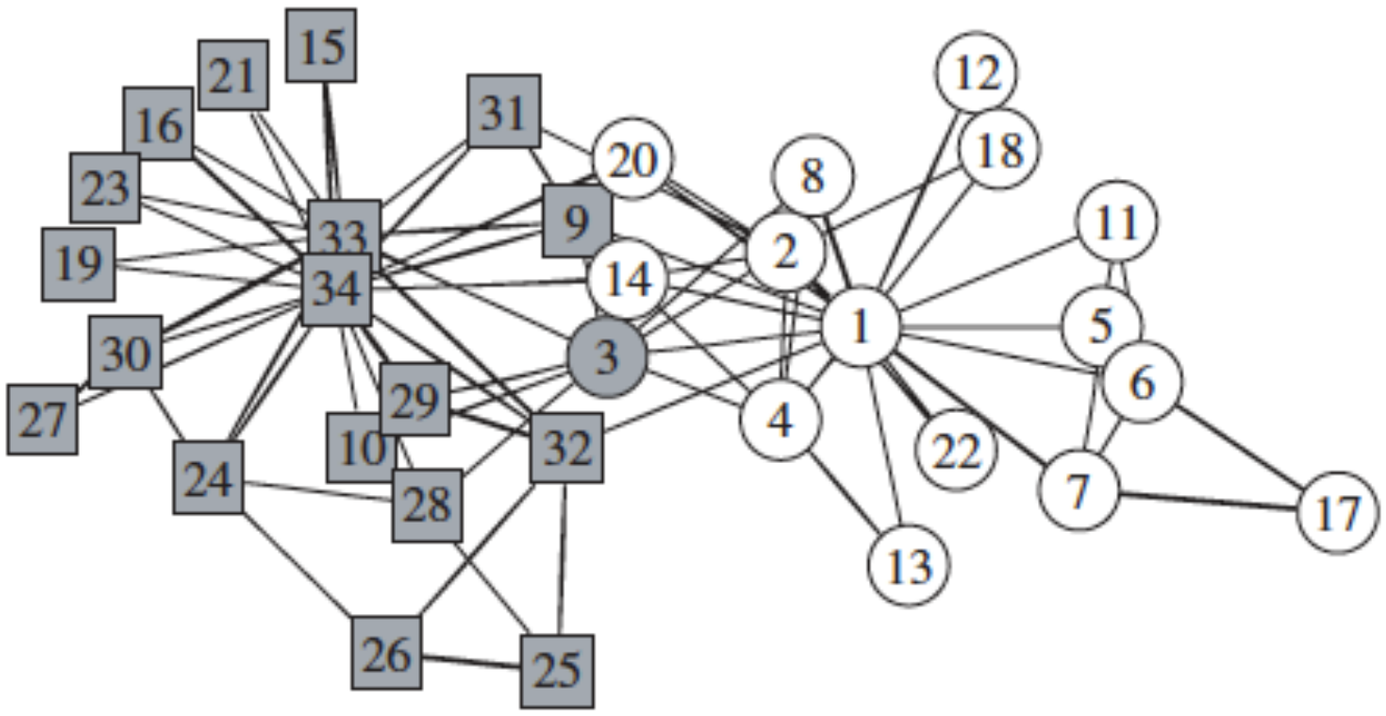
Example of social clusters: collaboration networks



Girvan & Newman, PNAS, 2002

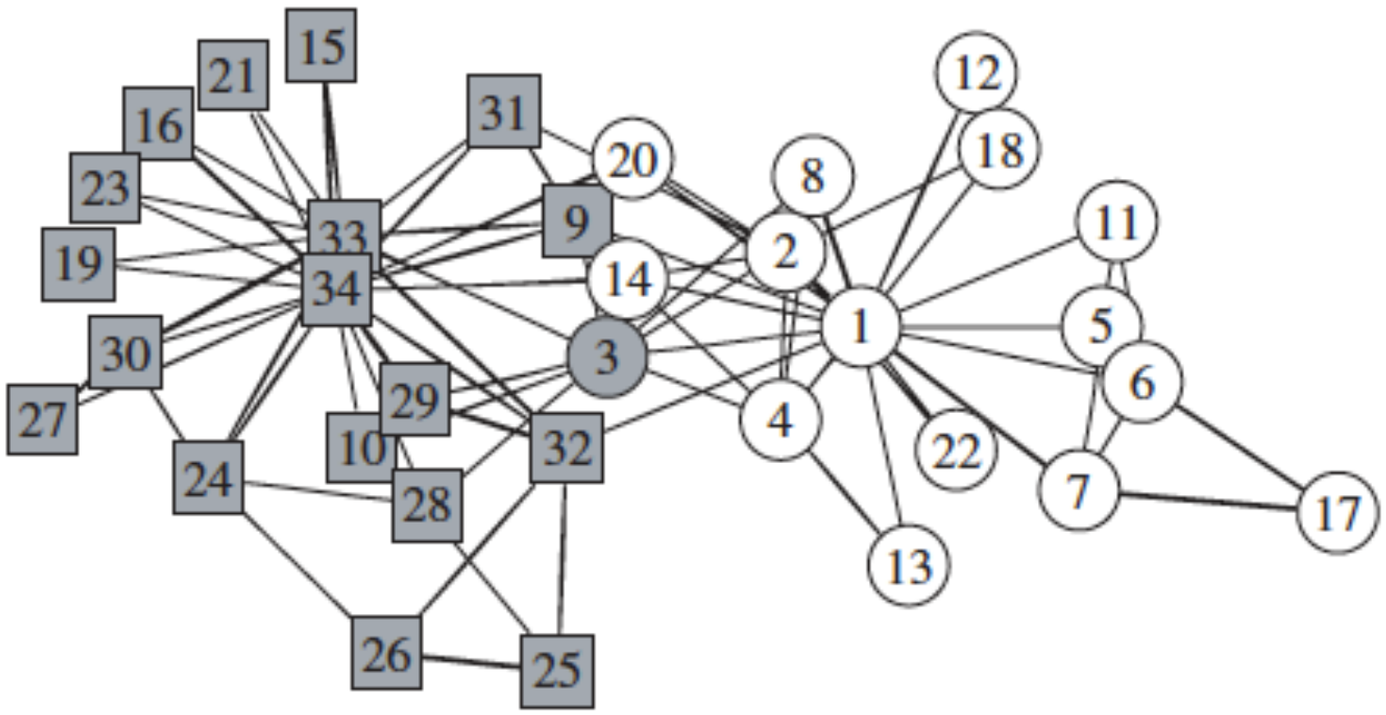
- Scientific collaboration network: link two scientists if they have co-authored a paper (indicating joint research work)
- Clusters in collaboration networks often have scientists working on similar topics

Example case: Zachary's Karate club network



- Social network of members of a karate club
- Disputes in the club, split into 2 parts
- Can the social network be used to predict which members joined which side?
  - Many clustering algorithms give accurate predictions for this network

Example case: Zachary's Karate club network



"If your method doesn't work on this network, then go home"

http://networkkarate.tumblr.com

Defining & detecting communities

- Target: divide the network into *partitions* or *covers*
  - Partition: each node belongs to exactly one cluster
  - Cover: each node can belong to any number of clusters (including 0)
- The definition can be an algorithm or description of how good community should look like

Two main approaches:

- Local definition:
  look at subgraphs and forget the rest of your network
- Global definition:
  find a partition/cover that is the best for the whole network

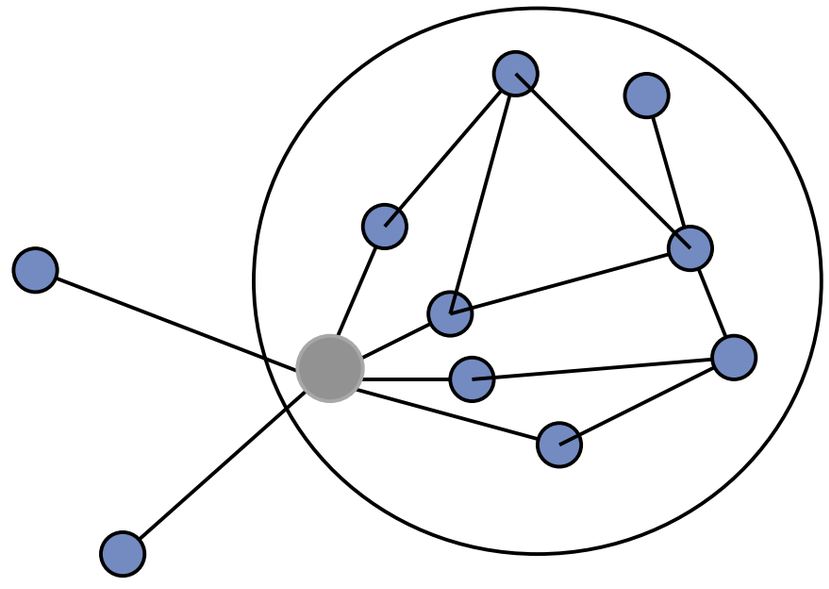
Local definitions: strong and weak communities

- Principle: comparison between internal and external cohesion of a subgraph

- **Strong community:** subgraph where the internal degree of each vertex is greater than its external degree (Luccio & Sami, 1969)

- **Problem**: very strict condition

- **Weak community:** subgraph whose internal degree is greater than its external degree (Radicchi et al., 2004)



Local definitions: clique percolation

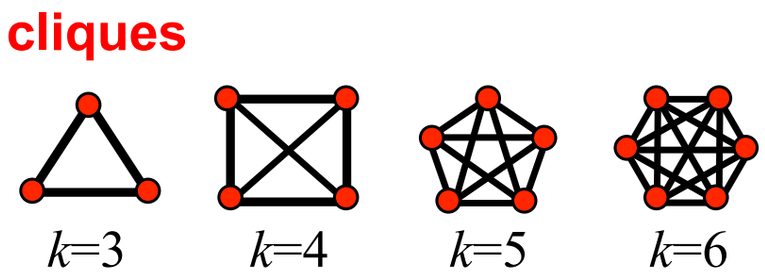
- Simple but appealing idea:

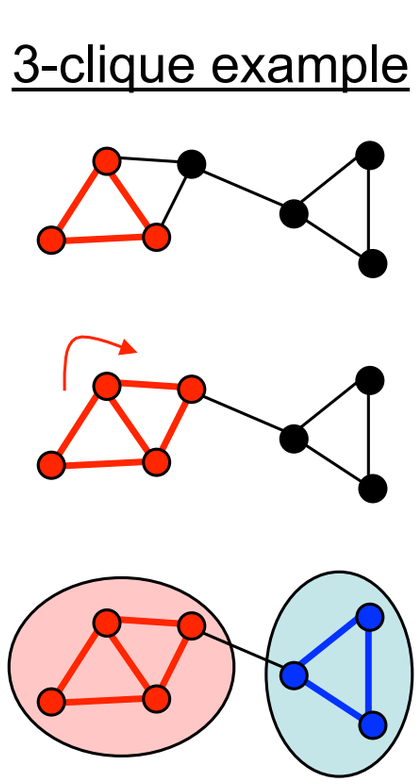
  a *k*-clique community is a set of nodes belonging to adjacent *k*-cliques, adjacent meaning that they share *k*-1 nodes

- Start from a clique and "roll it" until no more possible, label vertices, start from another & repeat

- Palla et al, Nature 435, 814 (2005)

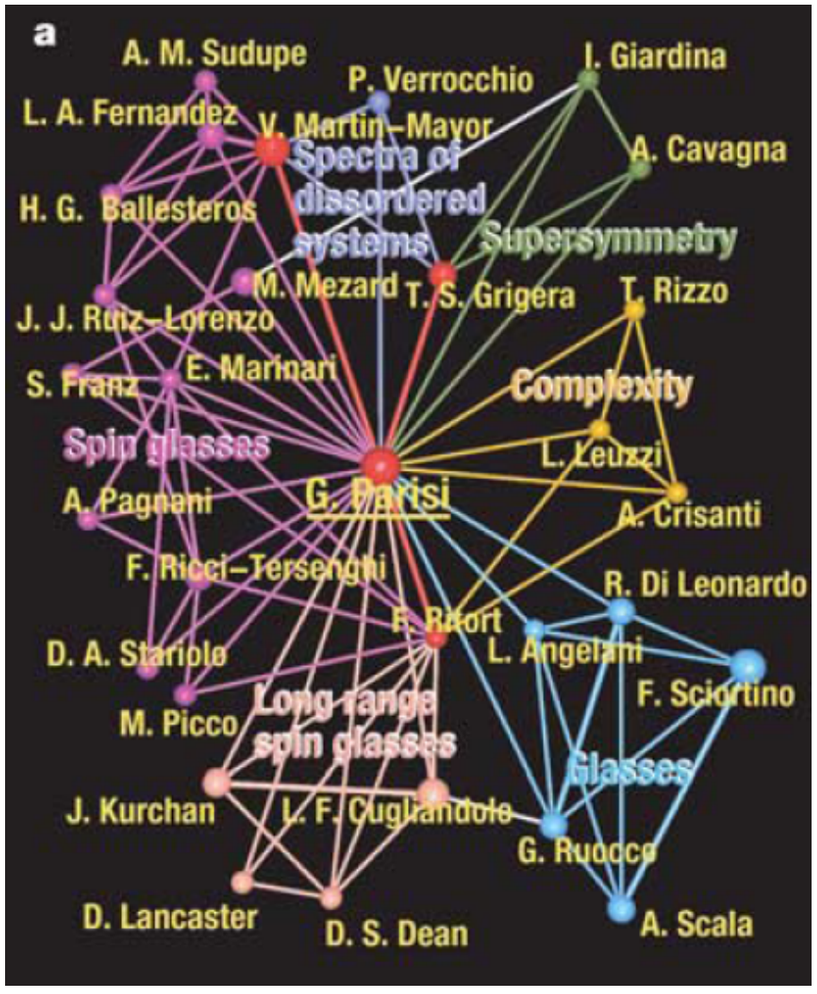
- Software available at http://www.cfinder.org/





Local definitions: clique percolation

- Finds covers (overlapping communities)



Palla et al, Nature 435, 814 (2005)

Global definitions: introduction

- Principle: **partition/cover assessed at the network level**

- Often, **quality functions** are used: Q(G,C) is a number determining how good partition/cover C is for a given network G

- Graph clustering is turned into optimisation problem:

$$\max_{C \in P(V)} Q(G, C)$$

  where P(V) is the set of all partitions of V

- Most interesting quality functions lead to a very difficult optimisation problems (often NP-hard)

- Therefore heuristic methods are used

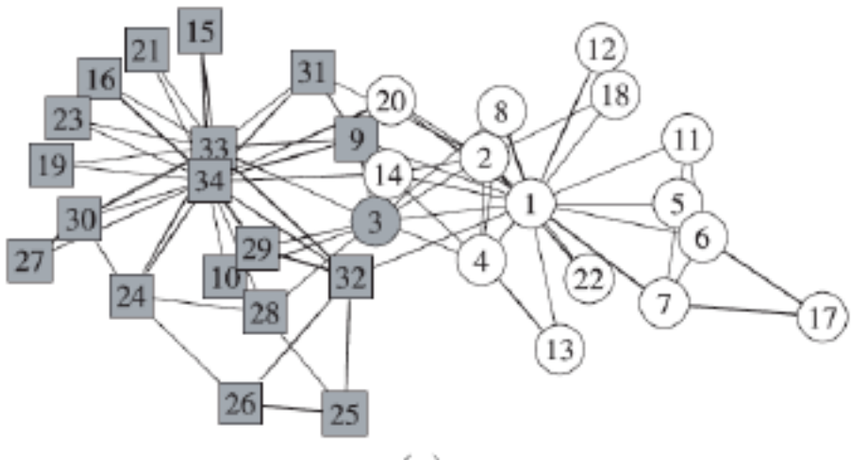
The Girvan-Newman method

1. Calculate the betweenness for all edges in the network.
2. Remove the edge with the highest betweenness.
3. Recalculate betweennesses for all edges affected by the removal.
4. Repeat from step 2 until no edges remain.

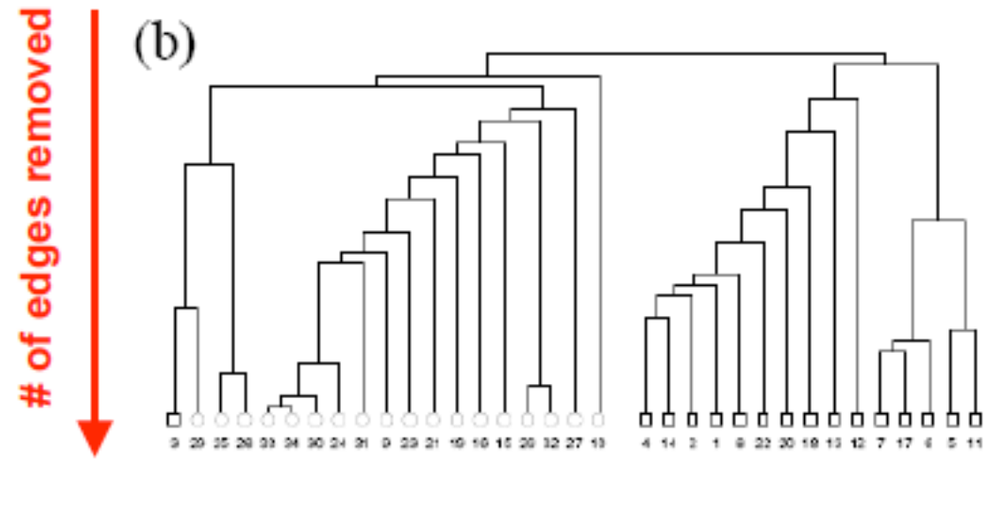
Girvan M. and Newman M. E. J., Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002)

- Edges connecting communities should have high betweenness centrality
- Idea: remove links in order of decreasing betweenness,
  say that communities = components in the remaining network
- So instead of one partition, a series of partitions, one for each step!

The Girvan-Newman method



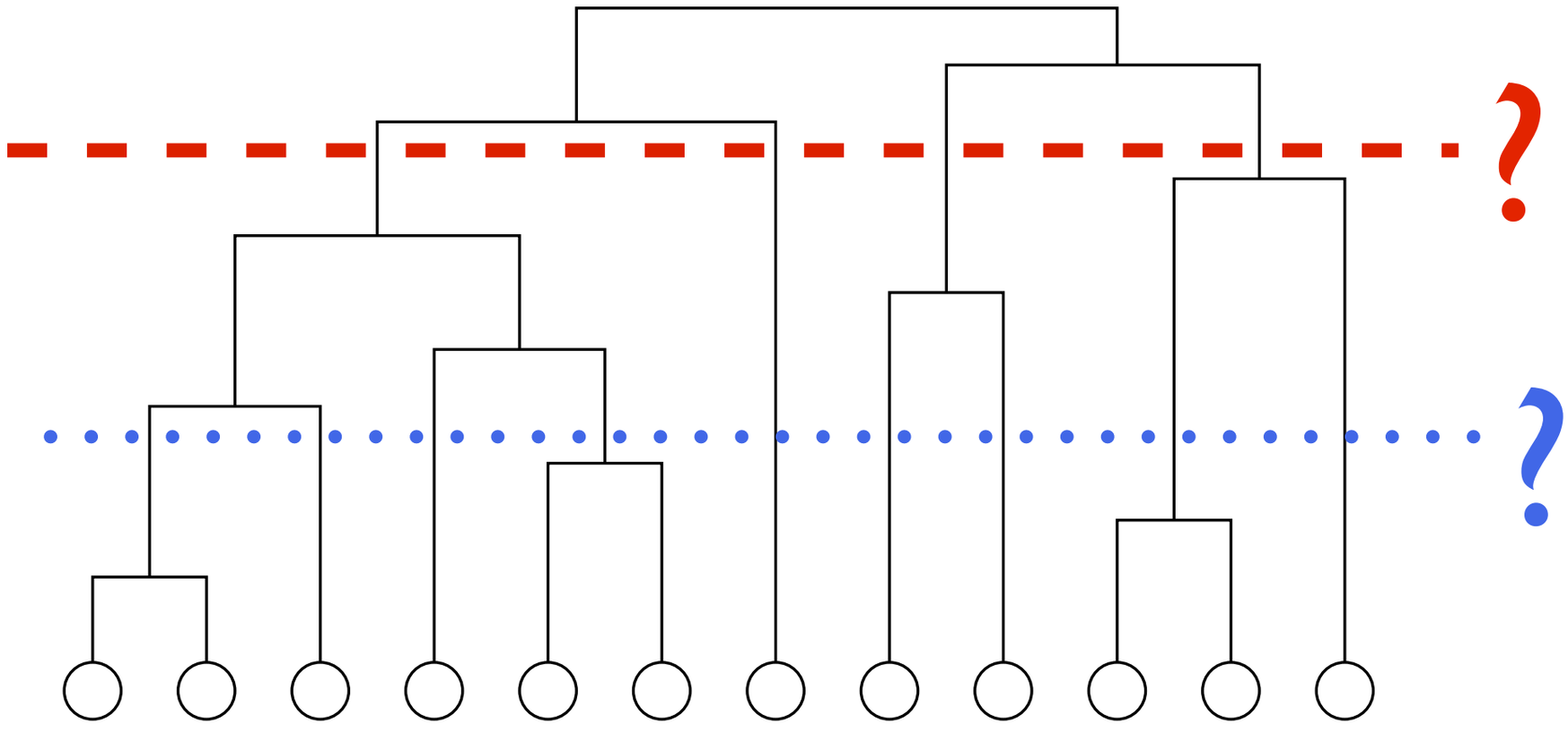
(a)



- Zachary's karate club, a famous example studied in social sciences
- The club split into two due to leaders' disputes
- First level of splitting with the GN method (circles, squares) corresponds to actual split

- This method is **hierarchical**, producing a tree structure of **nested** communities
- Such structures can be presented as **dendrograms**

Is there a best partition?



Modularity: which partition is the best one?

- The GN method produces many possible community divisions
- To select the "best" one, Newman proposed a quality function, the **modularity**

$$Q = \frac{1}{2E} \sum_{i \neq j} \left(a_{ij} - \frac{k_i k_j}{2E} \right) \delta(\sigma_i, \sigma_j)$$

counts only pairs of nodes in same community

this term measures the "excess" of community-internal links, when compared to reference model

probability of nodes i and j being connected in randomized reference networks (configuration model)

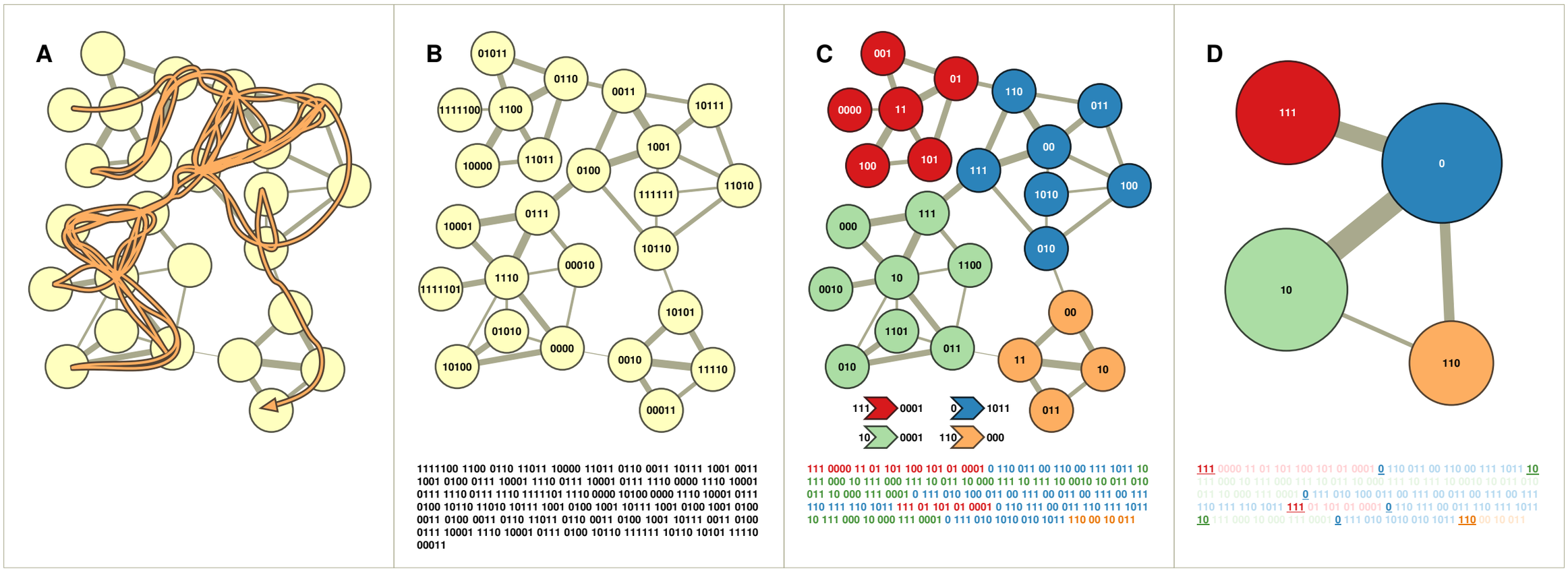
E =number of links in network,
a_{ij} = adjacency matrix element,
k_i = degree of i,
$\delta(x, y) = 1$ only if $x = y$,
σ_i = community label of node i

Modularity optimization

- Just use **any means to find a partition that optimizes Q!**
- There are many heuristics for this.
- The problem is **NP-hard**: difficult/impossible to find the global optimum, but reasonable solutions can be found.
- Problems:
  - **There is a resolution limit that depends on network size:** communities below a certain size are not detected (if they exist, they are merged together) (Fortunato & Barthelemy, PNAS 104, 2007)
  - **The method always produces communities,** whether there is anything resembling a community structure or not.

Model selection: INFOMAP

- Idea: give nodes & clusters Huffman codes, s.t. the description length of random walks would be minimized
- Random walkers: spend a lot of time inside a cluster, infrequently move between clusters



Maps of information flow reveal community structure in complex networks, Martin Rosvall and Carl T. Bergstrom, PNAS 105, 1118 (2008).

Free software: http://www.tp.umu.se/~rosvall/code.html

How to compare partitions?

- Jaccard index
- Mutual information
- Normalised mutual information
- Variation of information
- Rand index
- Omega index
- Precision/recall
- …

Information entropy

1. Give names to communities that are some sequences of binary numbers

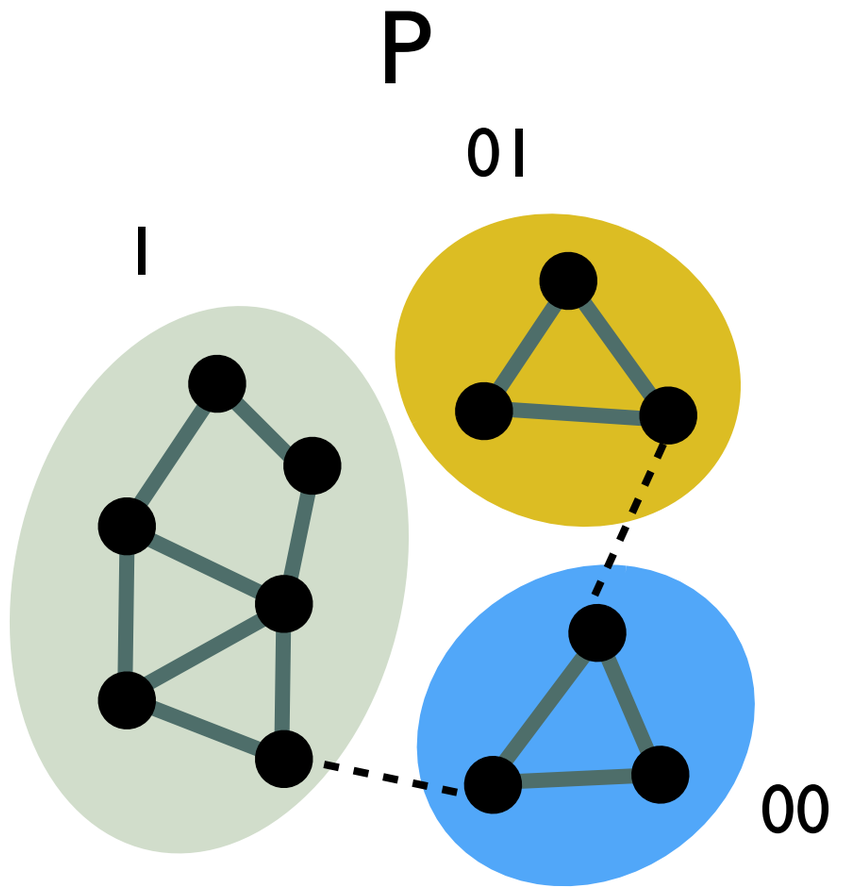
2. Pick a random node, the expected length of the name of the community the node belongs to depends on how good the naming is

3. Information entropy gives the minimum possible expected length:

$$H(P) = -\sum_{i=1}^{N_c} P(i) \log_2 P(i)$$

fraction of nodes in community C_i

$$P(i) = \frac{|C_i|}{N}$$



$$\begin{aligned} H(P) &= -0.5 * \log_2(0.5) - 0.25 * \log_2(0.25) \\ &\quad - 0.25 * \log_2(0.25) \\ &= 0.5 * 1 + 0.25 * 2 + 0.25 * 2 \\ &= 1.5 \text{ (bits)} \end{aligned}$$

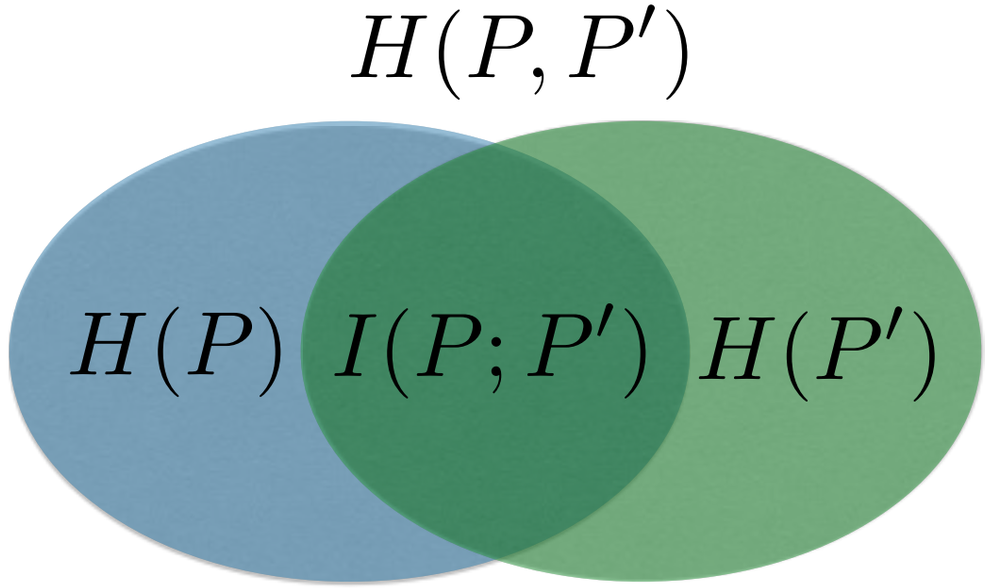
Joint information entropy

Instead of naming single communities in P one can give names to **all combinations** of communities in P and P'

$$H(P,P') = -\sum_{i=1}^{N_C}\sum_{i'=1}^{N'_C} P(i,i')\log_2 P(i,i')$$

probability of a random node to be in community i of partition P and community i' of partition P'

$$P(i,i') = \frac{|C_i \cap C'_{i'}|}{N}$$



Mutual information is the amount of information (entropy) that is shared between P and P':

$$I(P;P') = H(P) + H(P') - H(P,P')$$

If I know a nodes community in P how much information that gives about its community in P' (or vice versa)?

Normalised mutual information:

$$I_N(P,P') = \frac{2I(P,P')}{H(P)+H(P')}$$

Variation of information:

$$VI(P,P') = H(P) + H(P') - 2I(P,P')$$

Finding "the best" community detecting method

1. **Community detection methods** that detects a partition
2. **Benchmark model** that produces a network with a planted ground-truth partition
3. **Method for comparing** the ground-truth and the detected partition

Produce -> **detect** -> **compare**

Testing community detection methods

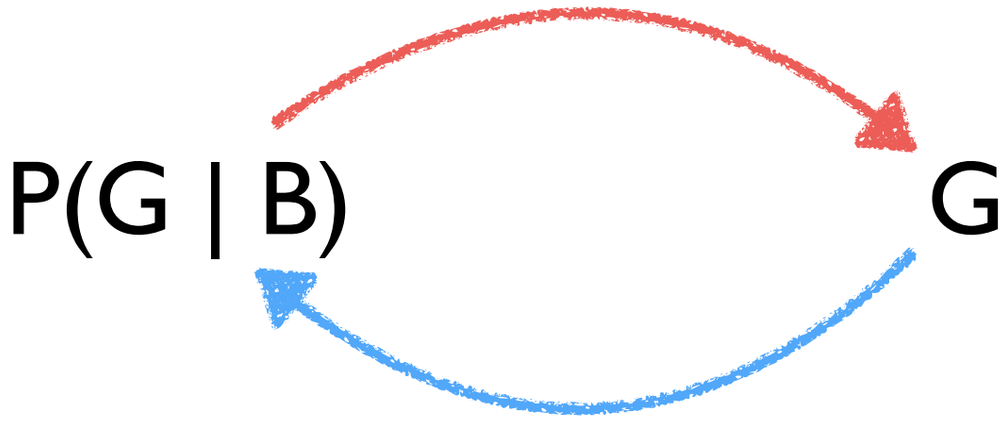
- Chicken-egg problem!
  - EVERY community detection method defines (often implicitly) what communities are
  - Then, it best detects communities that correspond to this definition
  - But we do not really know the community structure of a real network (and whether it corresponds to a particular definition)!

- Attempted solution: **benchmarks**
  - networks with **planted** community structure
  - however, the **benchmarks also implicitly define what communities are**
  - ....so only methods that agree with this definition can work well on them!

Stochastic block model

- A network model with structure can be used to detect structure via inference:

Generate network G with given blocks B



Find the blocks B that maximise the probability of producing the given G

- Stochastic block model:

$$P(G|B) = \prod_{(i,j)\in E} p_{B_i B_j} \prod_{(i,j)\notin E} (1 - p_{B_i B_j})$$

B_i : the block of node i

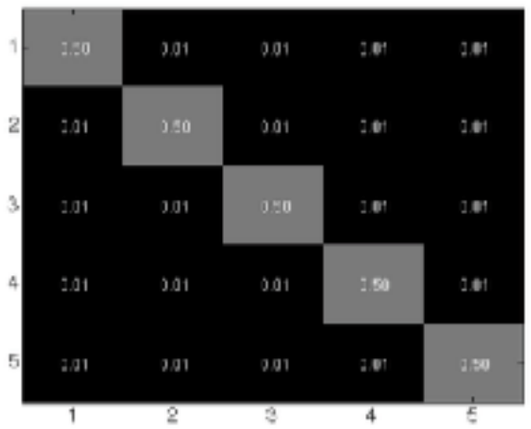
p_{BiBj} : prob of link between blocks B_i & B_j

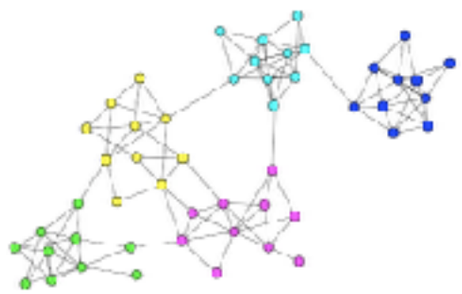
- One can use the theory of **statistical inference**

- In practice node degrees dominate the blocks: **degree corrected block models** are used

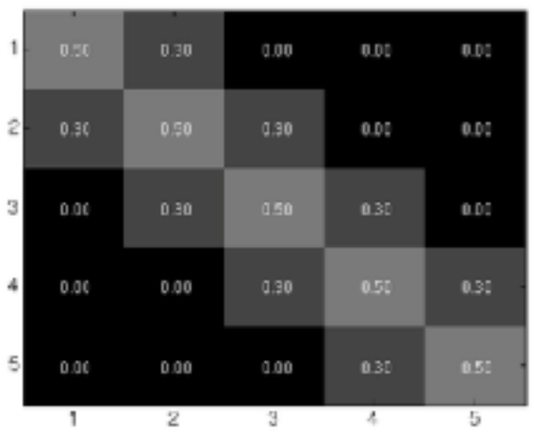
Stochastic block model

Communities

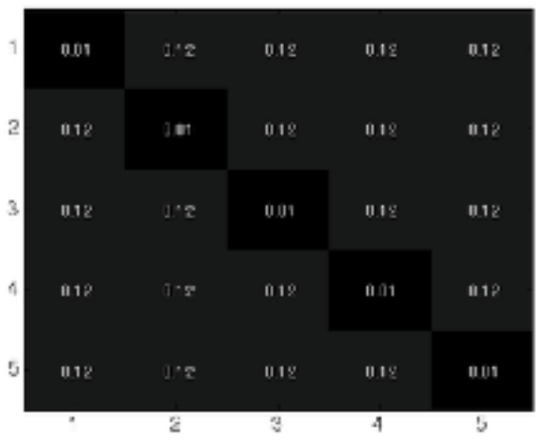


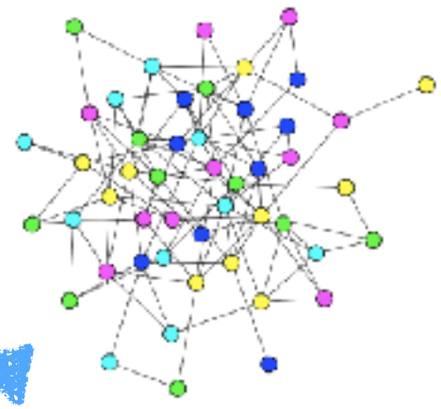


Ordered communities

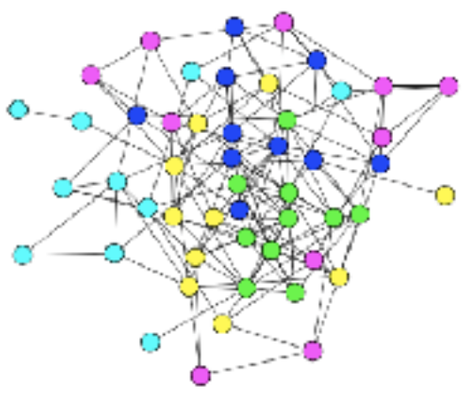


Disassortative blocks





Core-periphery structure



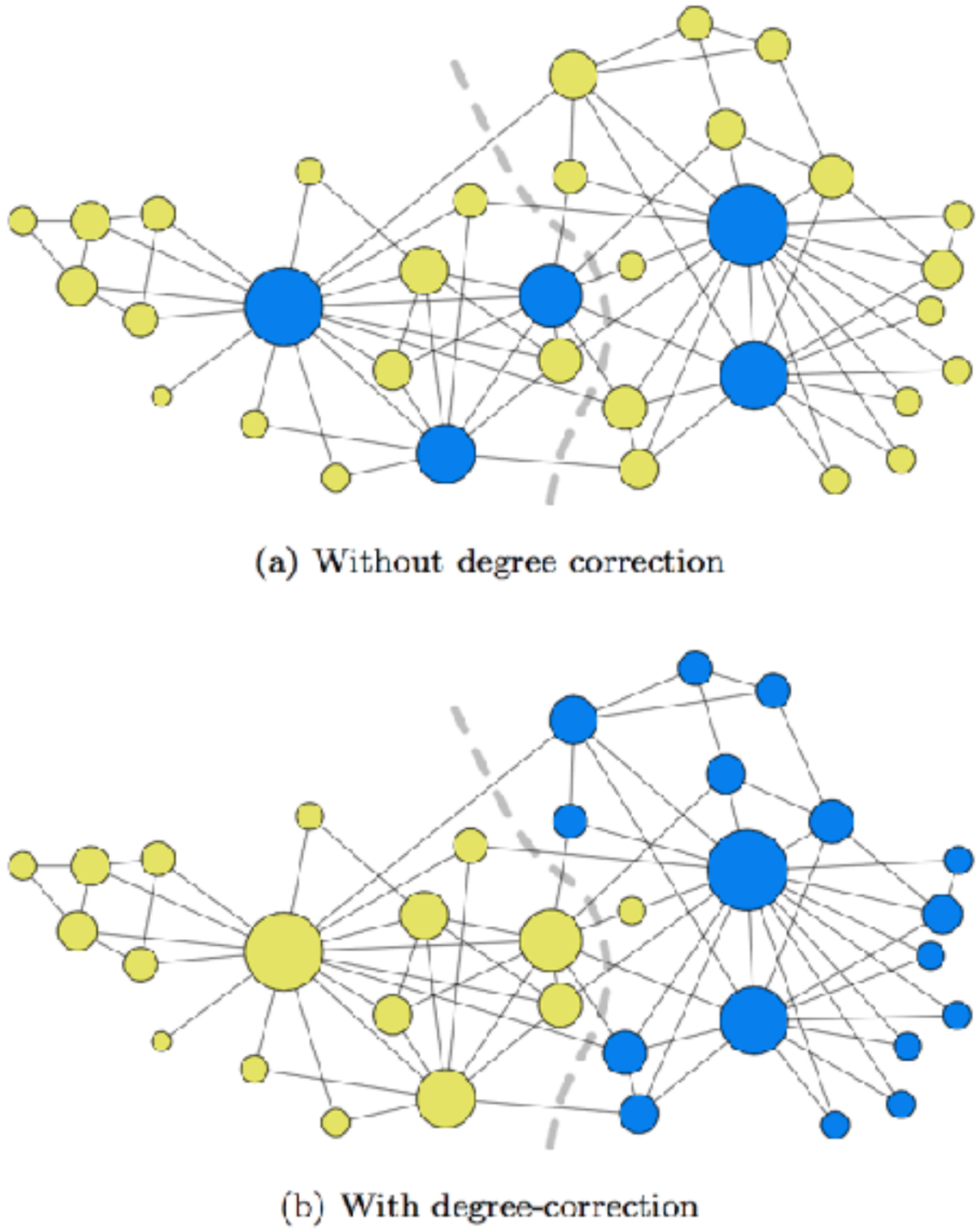
Sampled network G

Examples from:
http://tuvalu.santafe.edu/~aaronc/courses/5352/

Block matrix B with 5 blocks

Stochastic block model

- SBM fitting finds blocks that best explain deviations from totally random networks
- In practice variation in node degrees is often the most dominant structure
  - Degrees dominate the blocks: **degree corrected block models** can be used



(a) Without degree correction

(b) With degree-correction

Karrer & Newman, PRE 2011

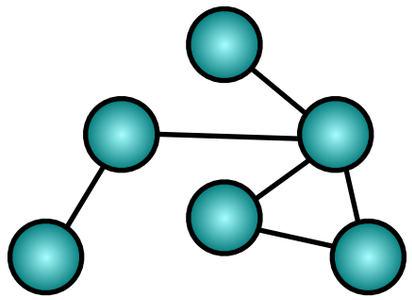
Conclusions & warnings

- there is NO correct solution to the community detection problem
- ALL methods define communities in different ways and detect subgraphs that match their definition
- there CANNOT be an universal method that suits all purposes
- win some = lose some

Sampling

What is a network sample?

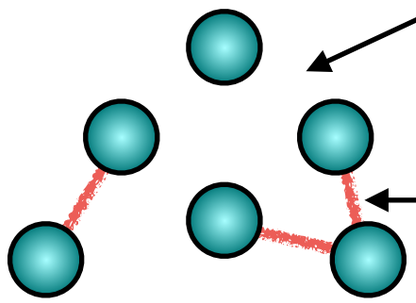
The original network



Sample edges

Edge missing as it is not in the sample

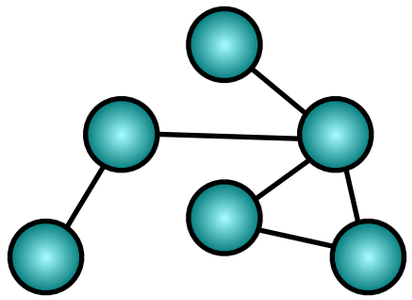
A sample network



Sampled edge

What is a network sample?

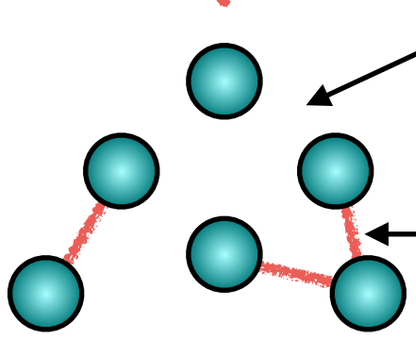
The original network



Sample edges

Edge missing as it is not in the sample

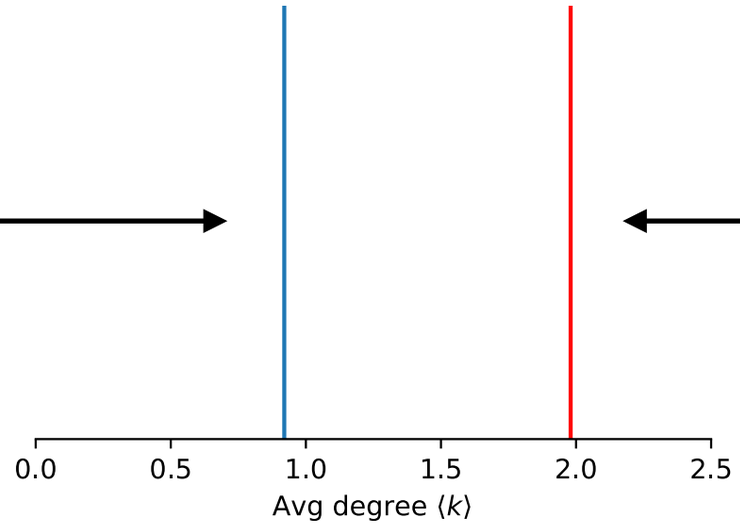
A sample network



Sampled edge

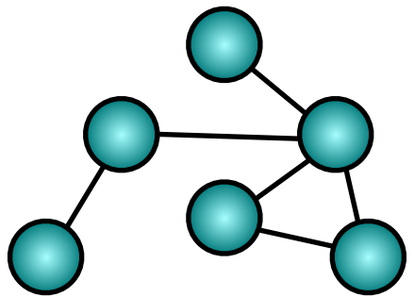
Average degree after sampling

Average degree in the original network (BA-model, n=100, m=1)



What is a network sample?

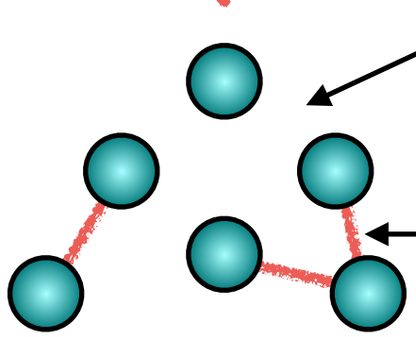
The original network



Sample edges

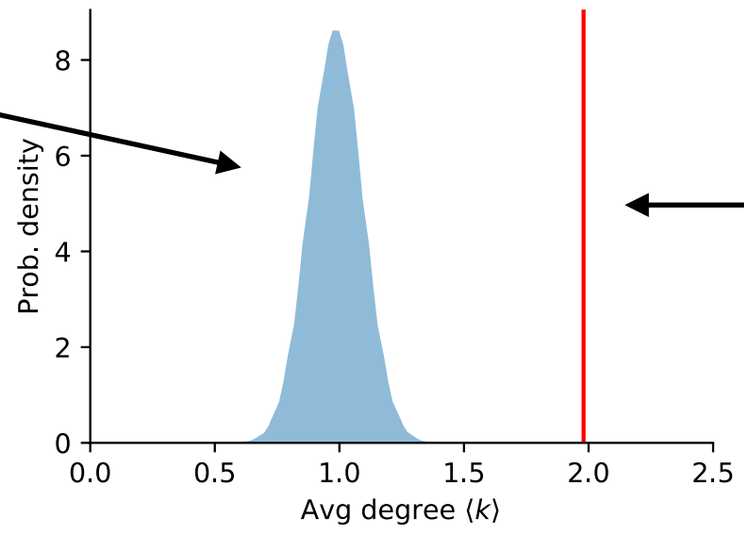
Edge missing as it is not in the sample

A sample network



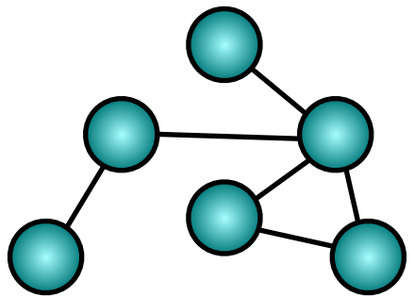
Sampled edge

Probability of average degree after sampling



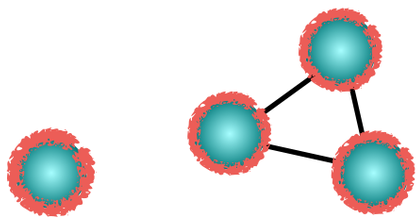
Average degree in the original network (BA-model, n=100, m=1)

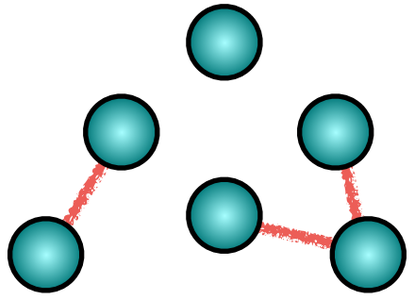
What is a network sample?

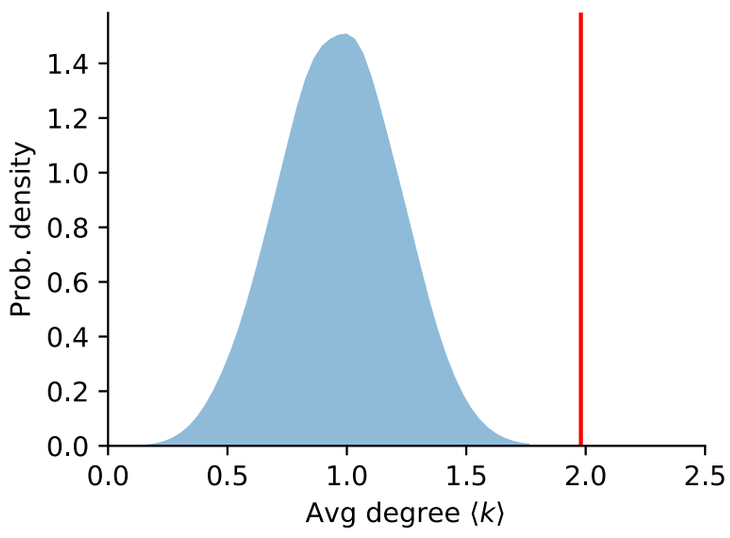


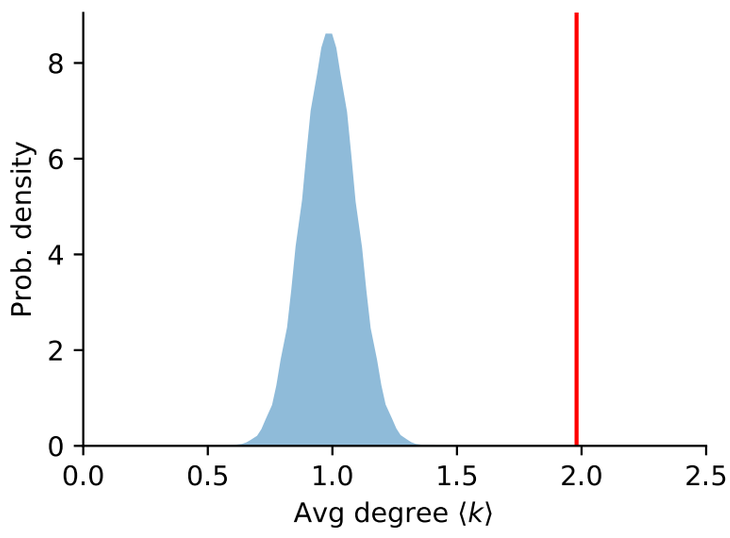
Sample nodes

Sample edges

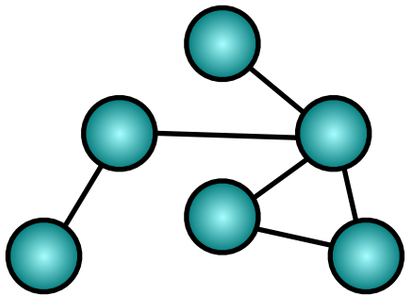








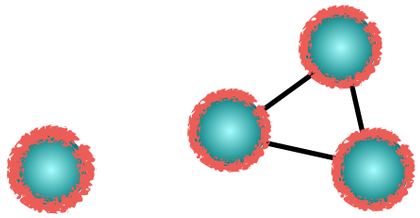
What is a network sample?

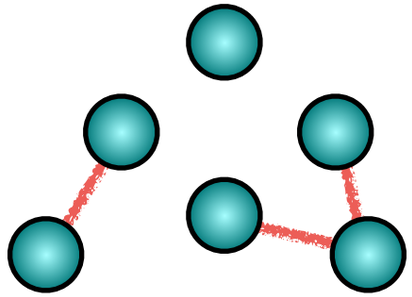


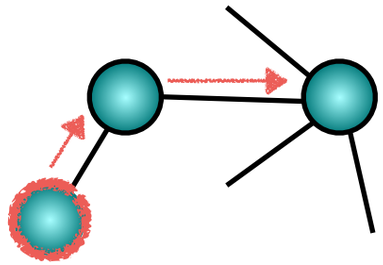
Sample nodes

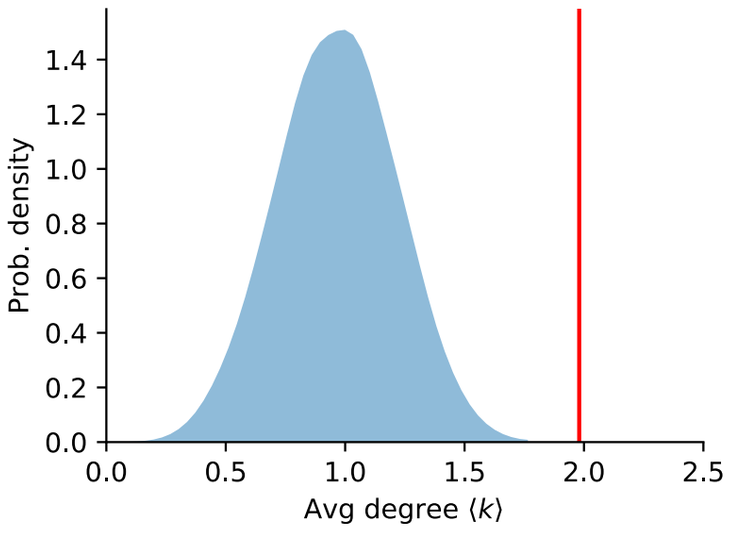
Sample edges

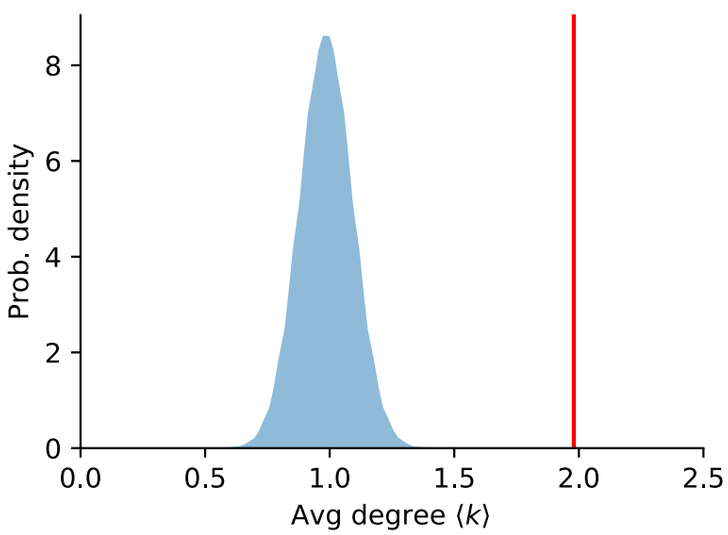
"snowball sample"

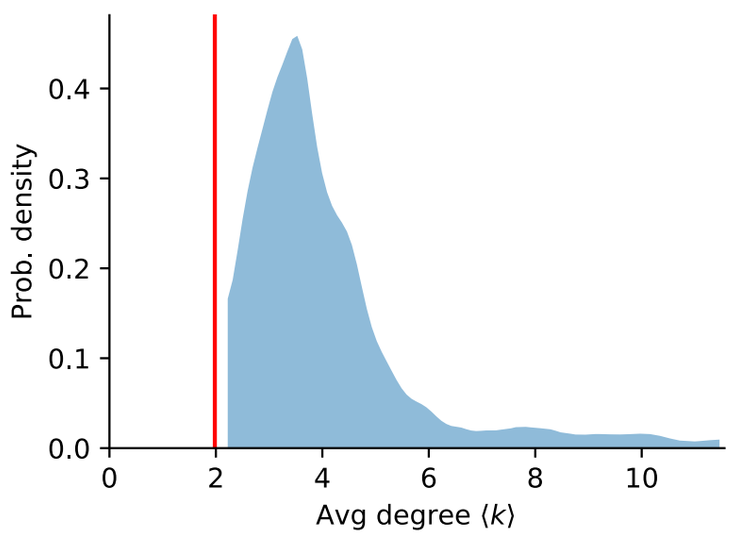












Most network data sets are samples

Most network data set introduced in these lectures are samples in some sense:

- **Mobile calling network** (lectures 6-7): Customers of the phone operator only 20% of population in the country, calls don't contain all social relationships
- **Internet autonomous systems**: Sampled with trace route queries between samples of source and target nodes
- **last.fm social network**: Sample of the "true" social relationships between the users of the website
- **Protein interaction network**: Some proteins are studied more than others, and their links are known better

Most (network) data sets are samples

Sampling data is very common, and there are good longstanding theories in statistic on how to deal with it:

- **Polls**: Election polls, opinion polls etc. are always samples of the full population. Errors can be estimated, biases corrected etc.
- **Blood tests**: Samples of the all the molecules in the blood.

Getting formal: How to deal with (network) samples?

- **Estimators:** Given a measured statistic and a sampling procedure, solve the inverse problem of finding the actual value of the statistic.

- Task: Given a sampled subset *S* of all elements *U*, estimate:

$$\tau = \sum_{i \in U} y_i$$

The value associated to sample i

- Horvitz-Thompson estimator:

$$\hat{\tau} = \sum_{i \in S} \frac{y_i}{\pi_i}$$

The probability to observe sample i

Example: average degree and edge sampling

- Out of all edges E sample each with probability p. Given the sample S, estimate the average degree:

$$k = \frac{1}{n} \sum_{u \in V} k_u$$

- … rewriting the formula noting that $k_u = \sum_{v \in N(u)} 1$:

$$k = \frac{1}{n} \sum_{u \in V} \sum_{v \in N(u)} 1 = \frac{2}{n} \sum_{(u,v) \in E} 1$$

- The sum and its estimator:

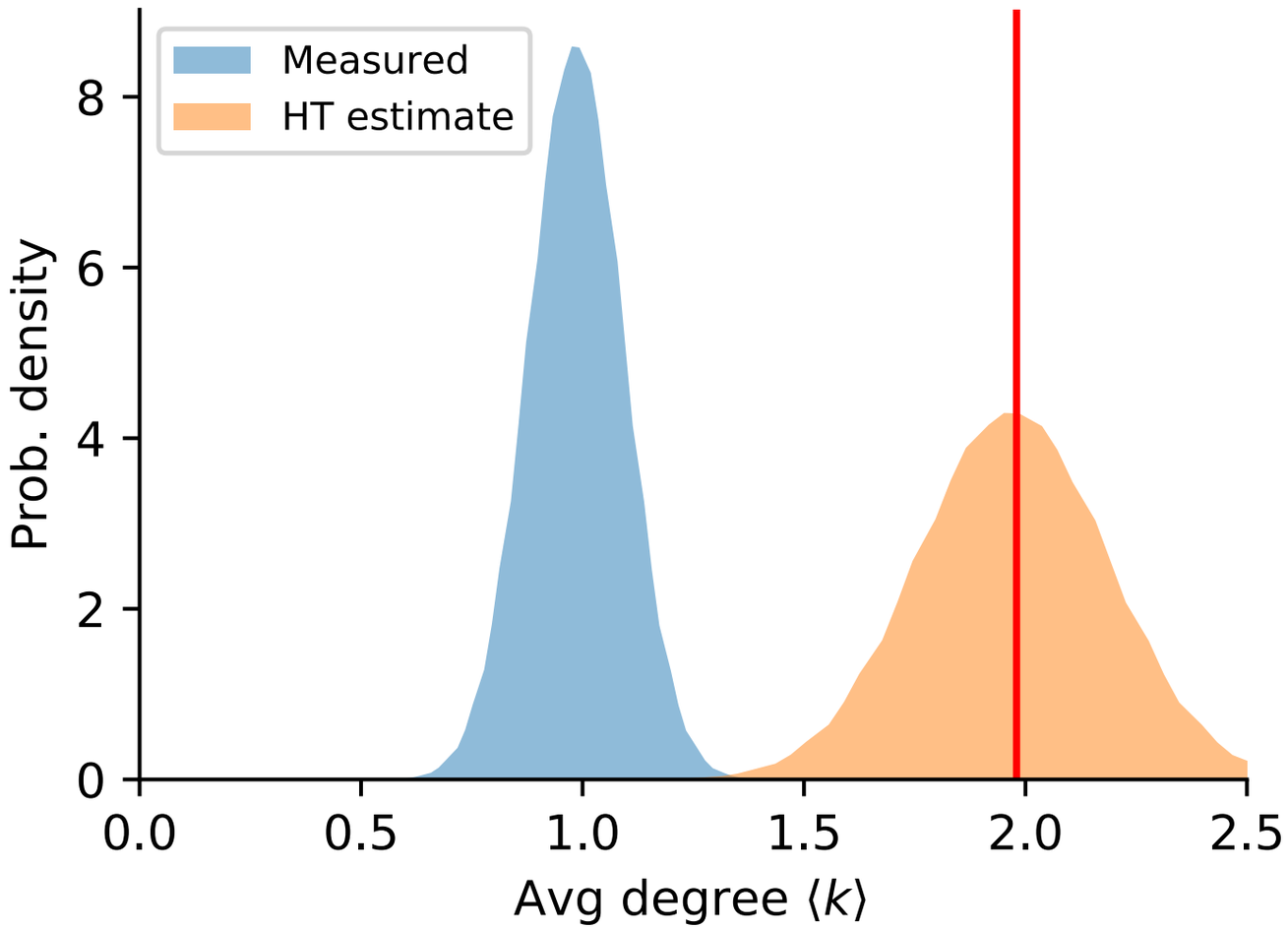
$$\tau = \sum_{(u,v) \in E} 1 \qquad \hat{\tau} = \sum_{(u,v) \in S} \frac{1}{\pi_{(u,v)}} = \sum_{(u,v) \in S} \frac{1}{p} = |S|/p$$

- Estimator for the average degree:

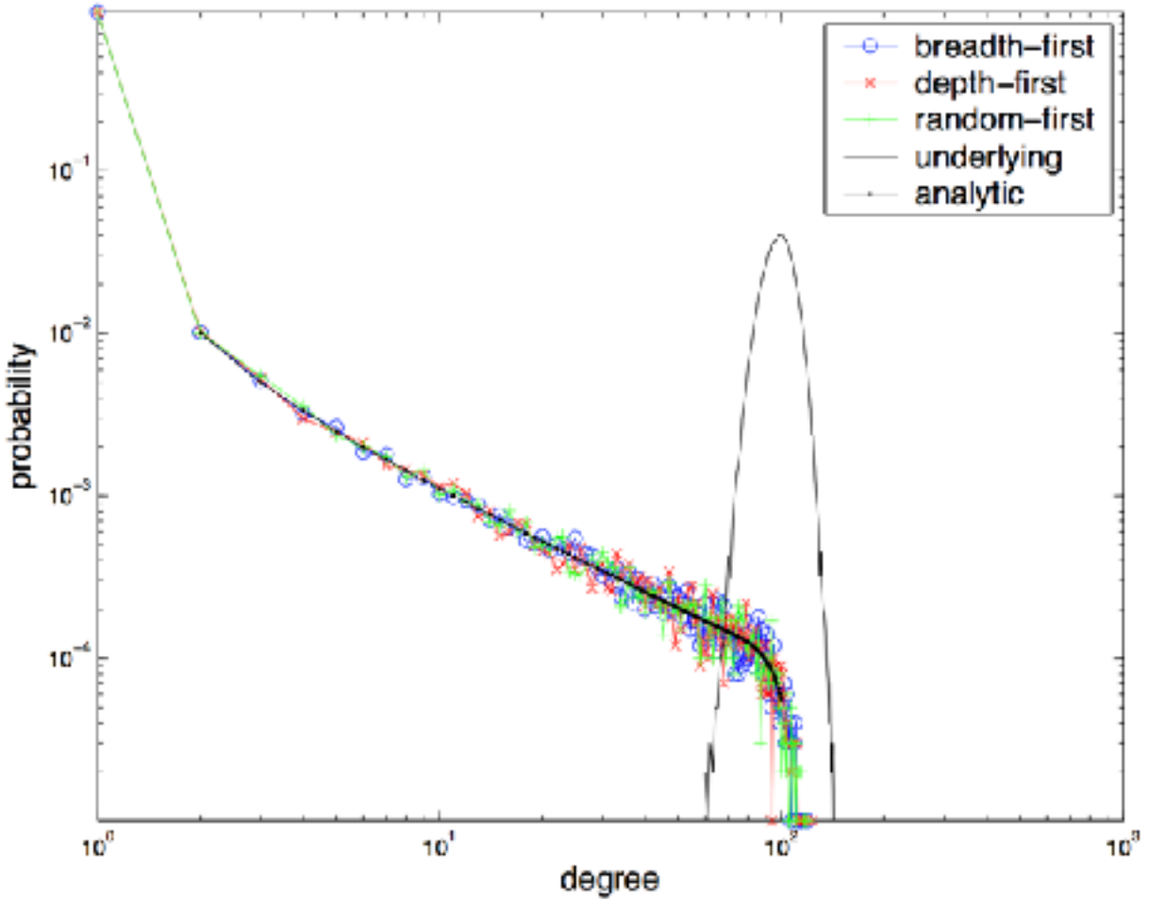
$$\hat{k} = \frac{2|S|}{np}$$

Example: average degree and edge sampling

- Estimator for the average degree: $\hat{k} = \frac{2|S|}{np}$



Can some results in the literature be there just because of biased sampling?



Clauset & Moore, PRL 2005

If you want more networks after this course

CS-E5745 - Mathematical Methods for Network Science

- Period III
- Learn the theory and mathematical tools behind central concepts in network science
- Lectures + exercises done with pen and paper (no exam)

CS-E5700 - Hands-on Network Analysis:

- Periods IV-V
- Learn to do network analysis in practise
- Analyse network data in projects done in small groups (no exams)